На правах рукописи

Алексеев Дмитрий Глебович

РАЗРАБОТКА АЛГОРИТМОВ ПРОТЕОГЕНОМНОГО ПРОФИЛИРОВАНИЯ МИКРООРГАНИЗМОВ

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва 2012

Работа выполнена в Учреждении Федерального медико-биологического агенства Научно-исследовательском институте физико-химической медицины (НИИ ФХМ ФБМА)

Научный руководитель	д.б.н., проф., чл.-корр. РАМН Говорун Вадим Маркович
Официальные оппоненты:	Лисица Андрей Валерьевич д.б.н., чл.-корр. РАМН, ФГБУ «ИБМХ» РАМН, зав.лаб.
	Николаев Евгений Николаевич д.ф-м.н., проф.,ФГБУН ИНЭПХФ РАН, зав.лаб.
Ведущая организация	МГУ имени М.В. Ломоносова, химический факультет

Защита состоится «12» апреля 2012 г. в 12:30 на заседании диссертационного совета Д 001.010.01 при Федеральном государственном бюджетном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н.Ореховича» Российской академии медицинских наук (ФГБУ «ИБМХ» РАМН) по адресу: 119121, Москва, ул. Погодинская, 10, стр.8.

С диссертацией можно ознакомиться в библиотеке ФГБУ «ИБМХ» РАМН

Автореферат разослан « 5 » марта 2012 г.

Ученый секретарь
диссертационного совета,
кандидат химических наук  Е.А.Карпова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Приближение к точке технологической сингулярности (Kurzweil 2005), охватившее сегодня все области человеческих знаний, во многом влияет и на исследования в области молекулярной биологии. Взрывообразное накопление данных в областях геномики, транскриптомики, протеомики и метаболомики не дает возможности перейти от редукционного подхода, направленного на отдельные компоненты, к системному, позволяющему охватить весь набор компонентов и их свойств.

Одной из отправных точек в исследовании живой системы является структура генома и его максимально полное описание – аннотация. Технологии, позволяющие получить геномную последовательность, получили повсеместное распространение и появилась возможность исследовать геном любого живого существа и даже отдельной клетки. При таких возможностях точность, полнота и скорость аннотации становится узким местом в исследованиях. Несмотря на большой арсенал развитых вычислительных методов создания геномных аннотаций, они принципиально являются лишь предсказательными.

Протеогеномика как предложенный в 2008 году набор подходов, основанных на использовании протеомных данных для улучшения геномной аннотации, позволяет существенно улучшить качество аннотации геномов. С учетом разнообразия царств Бактерий и Архей использование протеогеномной аннотации, возможно, является единственным способом получения корректного представления о связи генотипа и фенотипа. Было показано, что синтез наблюдений за относительно просто устроенными бактериальными клетками позволяет не только скорректировать представления о взаимоотношениях генов и их продуктов, но и получить представления о структуре системы в целом. Естественно, что появляющиеся в последнее время работы по созданию синтетической бактериальной клетки могут быть продолжены, только если создаваемая система будет полностью описана и смоделирована. Несмотря на актуальность и с учетом новизны названной области, сегодня не существует единого программного решения, которое бы объединяло все задачи , связанные с совместным использованием геномных и протеогеномных данных.

Цели исследования

Разработать подходы и алгоритмы протеогеномного профилирования бактериальных геномов, воплотить их в виде программного обеспечения и использовать для профилирования *Mycoplasma gallisepticum*, *Acholeplasma laidlawii*, *Spiroplasma melliferum*, *Desulfurococcus kamchatkensis* и *Helicobacter pylori*.

Задачи исследования

Для достижения названной цели были поставлены следующие задачи:

1) Разработка эффективных алгоритмов использования данных протеомных экспериментов для протеогеномного профилирования.

2) Использование алгоритмов для улучшения аннотации геномов *Mycoplasma gallisepticum*, *Acholeplasma laidlawii*, *Spiroplasma melliferum* и *Desulfurococcus kamchatkensis*.

3) Использование алгоритмов и оценка достоверности идентификаций при работе с изолятами и штаммами, для которых геномы не секвенированы или существует только частичная последовательность.

4) Использование алгоритмов для системного анализа и улучшения протеогеномной аннотации на основе сравнения протеогеномных профилей бактерий.

Научная новизна

С использованием современных методов и технологий разработаны оригинальные методики и алгоритмы обработки экспериментальных данных исследования геномов и протеомов бактерий. Комплекс подходов позволил впервые объединить в единое аналитическое пространство разрозненные данные частичного секвенирования ДНК и масс-спектрометрического анализа белков и далее, используя разработанный алгоритм протеогеномного сравнения, выявить межвидовые и межштаммовые различия.

Впервые проведено уточнение геномной аннотации для *Mycoplasma gallisepticum* S6, *Acholeplasma laidlawii* PG-8A, *Spiroplasma melliferum* KC-3 и *Desulfurococcus kamchatkensis* 1221n. По результатам уточнения удалось аннотировать новые белки, подтвердить или реаннторировать сайты начала транскрипции, проверить ряд предсказанных из строения генома явлений на белковом уровне. Ни для самих указанных штаммов, ни для близкородственных штаммов такие исследования ранее не проводились.

Проведенный с использованием разработанной методики анализ ряда бактерий позволил получить уникальные результаты по более точной оценке минимального функционального ядра молликут, исчерпывающему протеому представителя Архей, предположительным механизмам патогенеза насекомых у спироплазм и возможной особенности проявления вирулентности и способности к трансформации у бактерий вида Helicobacter pylori.

Практическая значимость.

Комплекс протеогеномного профилирования успешно используется в настоящее время в качестве основной информационной платформы в ряде международных и российских проектов, охватывающих как исследования бактериальной направленности (например, Метагеном и метапротеом микробиоты кишечника человека), так и исследования протеомов эукариот.

Предложенная методика протеогеномной аннотации, апробированная на различных представителях бактериального и архейного царств, может быть использована для протеогеномной аннотации любого бактериального или архейного генома. Предложенное использование системы с рядом дополнительных экспериментальных подходов (обогащение пептидной фракции протеома N-концевыми пептидами) позволит аннотировать большую часть экспрессируемых белков с точки зрения сайта начала транскрипции.

Система позволяет использовать данные современных экспериментальных установок с учетом их индивидуальных особенностей (точность, масштаб получаемых экспериментальных данных и т.п.). Использование любого

современного оборудования возможно за счет применения унифицированных стандартных форматов обмена данными.

Апробация работы.

Результаты работы были представлены на следующих российских и международных конференциях: *Итоговая научная конференция НИИ ФХМ 2010, Молодежная конференция НИИ ФХМ 2011 , BGRS 2010 Novosibirsk ,* XXII Симпозиум «Современная химическая физика» 2010 г. Туапсе, Постгеномные методы анализа в биологии, лабораторной и клинической медицины – 2010г. Москва, Постгеномные методы анализа в биологии, лабораторной и клинической медицины – 2011г. Новосибирск, *Iscb Students council 2011 – Vienna ,*HUPO 2011 World Congress – Geneve **,** MCCMB'11 -- Moscow .

Публикации. Материалы диссертационной работы отражены в 5 публикациях в рецензируемых российских и международных журналах и в 2 сборниках трудов конференций.

Структура и объем диссертации.

Диссертационная работа состоит из 4 глав (Обзор литературы, Материалы и методы, Результаты, Обсуждение), заключения и списка литературы содержащего 147 ссылок. Работа изложена на 105 страницах, содержит 25 рисунков и 5 таблиц.

Содержание работы.

1 Обзор литературы.

В обзоре литературы рассматривается основные методы и подходы, используемые при геномной аннотации и протеомной идентификации бактерий. Внимание уделяется уникальной ситуации, сложившейся в области бактериальной и архейной геномики – на сегодняшний день получено до 10 тысяч полных и частичных последовательностей разнообразных видов, при этом большая часть этих данных получена в последние годы (рис. 1).

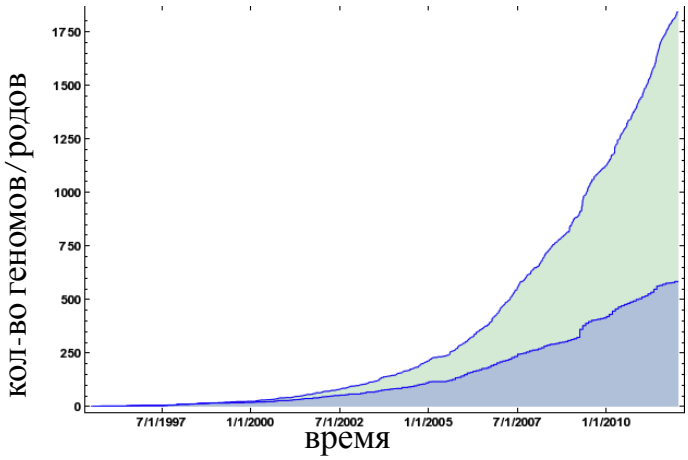

Рисунок 1 Экспоненциальный рост количеств геномов (зеленый) и родов (синий) со временем. График составлен по статистическим данным Genbank.

Анализируются основные подходы к геномной аннотации вычислительными методами – автоматизированные системы аннотации, которые используют *ab initio* предсказательные методики и методики, основанные на сравнении. Рассматриваются основные типы алгоритмов, предназначенных для идентификации белков по масс-спектру. Рассматриваются примеры оценки достоверности эвристических и вероятностных алгоритмов. На основе анализа литературы делается вывод о целесообразности использования нескольких алгоритмов(Brosch et al. 2008)(Kapp et al. 2005)(Colinge and Masselot 2004)(Nesvizhskii 2007) идентификации для достижения целей диссертационной работы.

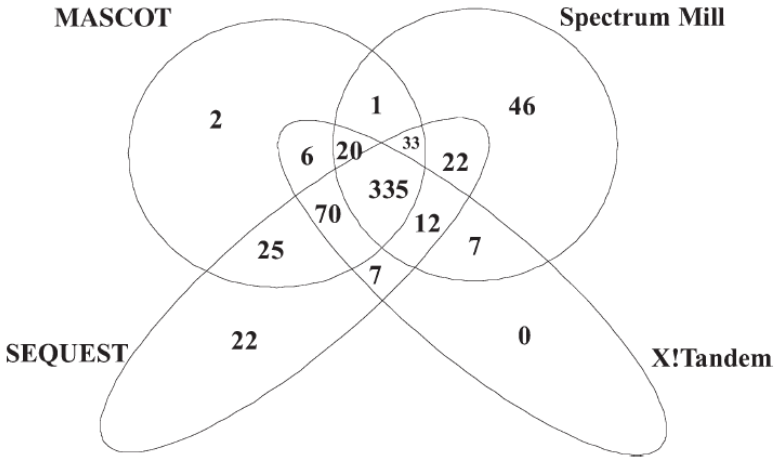

Рисунок 2 Диаграмма Венна отражающая количество пептидов идентифицированное наиболее распространенными алгоритмами.

В частности, выбор основывается на исследовании Броша и коллег(Brosch et al. 2008), где было показано, что алгоритм Mascot подходит для спектров снятых на ионных ловушках, в то время как алгоритм X!tandem показывает результаты лучше при использовании спектров с высокой точностью (менее 10 ppm).

В обзоре рассматривается несколько работ посвященных протеогеномному профилированию бактерий и архей. Перед проеогеномикой исследователями ставится ряд первоочередных задач:

1) Подтверждение предсказанных генов
2) Исправление предсказанных сайтов начала и окончания транскрипции
3) Идентификация не аннотированных генов

Рассматриваются примеры успешного применения технологии протеогеномного сравнения (сравнения, использующего похожести геномов для улучшения протеомной идентификации или построения системного анализа бактериальной клетки) на примере нескольких видов *Shewanella*(Gupta et al. 2008) или всего рода Микобактерий(Gallien et al. 2009).

Один из актуальных вопросов современной теоретической биологии – вопрос о минимальной клетке – или минимальном количестве генов, достаточных для существования автономно реплицирующейся формы жизни. В обзоре литературы

подробно разбираются методы и подходы позволяющие теоретически или экспериментально подойти к решению этого вопроса.

Приводится пример ряда работ последнего времени, расширивших горизонты представлений о сложности устройства бактериальной клетки. Так, было показано богатство некодирующих РНК в работах с транскриптомами *Bacillus subtillis* (Rasmussen, Nielsen, and Jarmer 2009)*, Mycoplasma pneumonia* (Güell et al. 2009) и *Helicobacter pylori* (Sharma et al. 2010), при этом наблюдается тренд: при совершенствовании технологий профилирования количество наблюдаемых некодирующих РНК растет от 209 для технологии РНК-чипов до почти 1000 при использовании глубокого сиквенирования, при этом возможно, что при приближении экспериментальных технологий к изучению единичной клетки количество наблюдаемых РНК, не кодирующих белковые продукты, может превысить количество генов. Кроме того, рассматриваются работы по идентификции белкового комплексообразования, где, на примере *M. Pneumonia* (Kühner et al. 2009), удалось обнаружить не только обилие комплексов, охватывающих почти все основные функциональные белки клетки, но и неожиданные комплексы, в которых взаимодействовали белки метаболизма и белки синтеза белка. Такие факты наводят исследователей на мысль, что количество контактов между компонентами реальной клетки выходит за рамки известных функциональных. Более того в работе по *Bacillus subtillis* (Commichau et al. 2007) с использованием дву-гибридной дрожжевой системы подтверждено взаимодействие между ферментами гликолиза енолазой и фос-фруктокиназой и ферментами, участвующими в процессинге РНК, являющимися жизненно важными. Такая находка заставляет авторов высказать предположение о том, что повсеместное присутствие генов гликолиза связанно именно с взаимодействием с жизненно важными генами и возможной транскрипционной модуляцией активностей участников контакта в таких комплексах, и, кроме всего прочего, структурной функцией.

Кроме того, в обзоре литературы представлено краткое описание исследуемых видов бактерий: нескольких миколпазм, *Desulfurococcus kamchatkensis* и *Helicobacter pylori*.

2 Методы

В данном разделе рассматриваются методы, технологии и алгоритмы, использованные для создания платформы протеогеномного профилирования микроорганизмов.

2.1 Создание экспериментальной базы данных

Все экспериментальные данные были размещены в реляционной базе данных, основанной на СУБД Oracle 11g. Настройка связей при нормализации и выделение ключевых сущностей при построении структур базы данных позволяет удерживать все экспериментальные данные в едином информационном поле и совместно использовать разнородные экспериментальные данные. Кроме того, для ускорения взаимодействия с публичными хранилищами данных были созданы усеченные реплики данных в экспериментальной базе данных. Подобное решение позволяет производить сравнения любых наборов экспериментальных данных с использованием простых SQL запросов. В процессе работы часть таблиц была денормализована для повышения производительности. Общая схема БД содержит более 200 таблиц, на схеме приведены основные сущности.

2.2 Программные пакеты для протеомного анализа и параметры обработки спектров.

Для первичной обработки спектров использовались пакеты Bruker Data Analysis, Agilent Mass Hunter. Полученные данные о спектрах развала были экспортированы в формат Mascot generic для последующей обработки.

Для идентификации использовался пакет идентификации Mascot 2.1.04 и пакет X!Tandem release 2008.02.11. Для обоих пакетов в случае работы со спектрами ионных ловушек использовалась точность 0,5 Да для родительского иона и 0,5 Да для спектра распада, в случае работы со спектрами Q-TOF точность 5 ppm для родительского иона и 0,5 Да для спектра распада.

2.3 Программы для сравнения геномов и картирования ридов

Для выравнивания геномов и контигов с последующим обнаружением полиморфизмов использовался пакет Mummer 3.0(Delcher et al. 2002).

Для выравнивания геномных ридов на геномы использовался пакет Bowtie build 0.12.5(Langmead et al. 2009). Обнаружение полиморфизмов производилось при помощи пакета SAM tools(Li et al. 2009).

Обнаружение аминокислотных полиморфизмов и создание на основе данных белковых баз данных для поиска производилось с помощью собственного ПО, реализованного в виде веб-сервиса с использованием технологий ASP.NETи extJS.

2.4 Объединение сторонних программ в Автоматизированный программный конвейер

Объединение программ в конвейер производилось на основе принципов построения распределенных систем с использованием веб-серверных технологий apache (для платформ на основе ядра linux) и IIS (для платформ на основе windows). Очереди задач с отсроченным исполнением по мере освобождения ресурсов создавались с применением технологии системных служб. Обработка форматов производилась с использование языка скриптов perl и C#.

2.5 Разработка пользовательских интерфейсов

Пользовательский веб-интерфейс разработан для наиболее часто используемых функций (регистрация экспериментов, просмотр результатов протеомных поисков, объединение в проекты и протеомное сравнение, сравнение геномов, картирование ридов и проч.) на основе технологий ASP.NETи extJS.

2.6 Статистический анализ

Статистический анализ производился с использованием пакета R(Ihaka and Gentleman 2007) version 2.14.1, для удобства использовалась надстройка RStudio v0.95.

Многофакторный анализ использования кодонов производился с использованием пакета CodonW (“http://codonw.sourceforge.net” n.d.)

2.7 Программирование алгоритмов

Алгоритмы были запрограммированы с использованием встроенного языка анализа данных pl/sql внутри экспериментальной базы данных. Подход был выбран в связи с высокой эффективностью индексирования больших объемов данных в промышленных СУБД, при этом вычислительная сложность алгоритмов была

невысокой, что позволило реализовать алгоритмы внутри БД без ущерба для общей производительности.

2.8 Протеомные эксперименты

Для получения исчерпывающего протеома нами (Совместно с Деминой И.А.) была разработана следующая схема фракционирования, примененная во всех экспериментах.

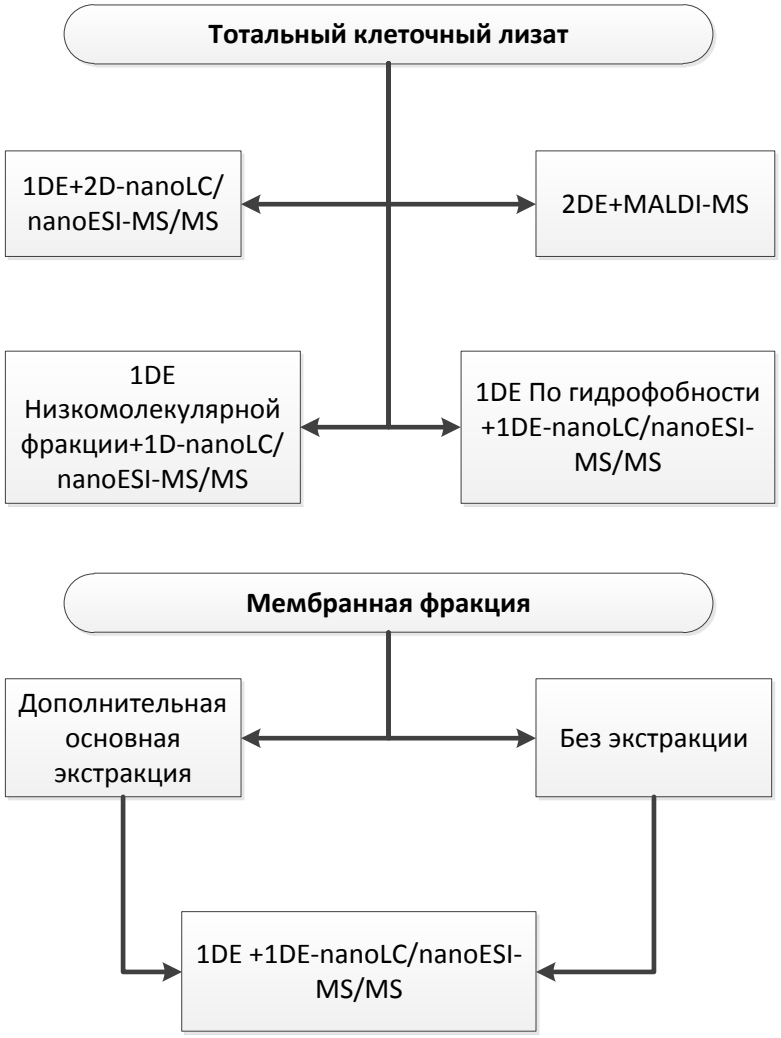

Рисунок 3 Схема фракционирования белковых смесей примененная в протеомных экспериментах для получения исчерпывающего протеома

Описание анализа приведено подробно в работе по *S.melliferum* (Alexeev et al. 2012).

2.9 Получение культур клеток

Культуры клеоток было получены в лаборатории протеомного анализа НИИ ФХМ, в кратце:

Культуру клеток *M. gallisepticum* S6 выращивали на жидкой среде, содержащей 2% триптозы, 0.5% глюкозы, 0.5% NaCl, 0.13% KCl, 0.3–0.5% Трис, 5% (v/v) дрожжевого диализата, 10% (v/v) сыворотки лошади, pH 7.2, при 37оС в течение 18 ч. *A. laidlawii* PG-8A выращивалась на жидкой модифицированной среде Эдвардса при 37оС в течение 18 ч.Для мониторинга роста культур использовался краситель феноловый красный.

Культура клеток. *S. melliferum* KC3была получена от профессора Вроблески (Université de Rennes, France). *S. melliferum* KC3 была выращена на среде SP4 как описано ранее(Tully et al. 1977). Сбор культуры проводился в log фазе, контроль производился по pH (7.2-7.0).

Культура клеток *D. kamchatkensis* была получена от проф. Бонч-Осмоловской (Институт микробиологии им. С. Н. Виноградского РАН, Москва)

Культуры клеток *H. pylori* были выращены на среде с кровяным агаром 2 – 3 дня в микроаэробном окружении (5% O_2, 10% CO_2, и 85% N_2) при 37°C.

2.10 Геномные эксперименты

Эксперименты по геномному секвенированию были проведены в лаборатории Постегномных методов исследований НИИ ФХМ, в кратце применялись следующие методики:

Секвенирование по Сангеру. Библиотеки заданной длинны (2 kb) были проведены через клонирование с вектором pCR4Blunt-Topo, далее выращены в *Escherichia coli* TOP10 и секвенированы при помощи BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, USA).

Секвенирование на SOLiD. Получали библиотеки спаренных фрагментов для SOLiD™ 4 system (Applied Biosystems, USA) со вставками примерно 2.5 КБ или 5.5 КБ. Для библиотек получали фрагменты длиной 50 нуклеотидов с F3 и R3 метками на концах.

2.10.1 Методы сборки

Для длинных геномных прочтений, полученных по методу Сангера, сборка геномов производилась с использованием пакетов: Phred(Ewing et al. 1998) – для

отбора качественных прочтений, LUCY(Chou and Holmes 2001) – для удаления последовательностей векторов, TIGR Assembler(Sutton et al. 1995) – для непосредственной сборки, и BAMBUS(Pop, Kosack, and Salzberg 2004) – для создания скаффолдов по данным о парности ридов в библиотеке.

2.10.2 Аннотация

Исходный набор ОРС возможно кодирующих белки определялся пакетом Artemis(Rutherford et al. 2000). Предсказанные ОРС длиннее чем 100 а.о. были использованы для поиска гомологов с помощью программы BLASTP(Altschul et al. 1990) в неизбыточном наборе генов NCBI и далее были вручную аннотированы с использованием информации о гомологии. Для аннотации использовался пакет визуализации UGENE("Unipro UGENE: an open-source bioinformatics toolkit; http://ugene.unipro.ru" n.d.). Ортологи были определены с использованием критерия взаимного наилучшего совпадения(G. S. A. Myers et al. 2006). Начала трансляции были определены из выравниваний BLASTP. Серверы TMHMM(Krogh et al. 2001) и HMMTOP(Tusnády and Simon 2001) были использованы для определения трансмембранных доменов. Рибосомальная и транспортные РНК были идентифицированы с помощью BLASTN(Altschul et al. 1990) и tRNA-Scan-SE(Lowe and Eddy 1997), соответственно. Метаболическая реконструкция производилась с использованием KEGG("(http://www.genome.jp/kegg/pathway.html)" n.d.).

2.11 Источники геномных данных

Для работы с геномами были использованы следующие версии геномов, доступные в NCBI.

Таблица 2. Использованные в работе версии геномов.

Организм	Версия генома в NCBI
Mycoplasma gallisepticum str. R(low)	CP001872.1
Mycoplasma gallisepticum str. F	CP001873.1
Mycoplasma gallisepticum str. R(high)	NC_004829.2
Helicobacter pylori 26695	NC_000915.1
Helicobacter pylori J99	NC_000921.1
Mycoplasma mobile 163K	NC_006908.1

Desulfurococcus kamchatkensis 1221n	NC_011766.1

Геномы, сиквенированные и аннотированные в ходе работы, были размещены в Genbank под следующими идентификаторами

Таблица 3. Идентификаторы размещенных геномных данных.

Организм	Идентификатор в Genbank
Acholeplasma laidlawii strain PG-8A	NC_010163.1
Mycoplasma gallisepticum S6	AFFR00000000.1
Spiroplasma melliferum KC-3	AGBZ01000000

Для работы с геномом *S.citri* GII3-3X была использована версия генома, полученная Карле и коллегами(Carle et al. 2010) , на сегодняшний день она доступна только на сайте лаборатории авторов(“Spiroplasma citri genome” n.d.)

3 Результаты и обсуждение

Нами был разработан набор оригинальных алгоритмов и автоматизированный программный конвейер (АПК) для обработки данных протеомных экспериментов, позволяющий решать полный цикл задач протеогеномного профилирования. АПК позволяет использовать данные протеомных и геномных экспериментов и получать на выходе биологические выводы в виде таблиц, графиков и диаграмм.

АПК включает в себя набор программных средств, созданных другими исследователями для обработки данных, набор собственных алгоритмов, разработанных нами на основе методов, изложенных в обзоре литературы, а также несколько алгоритмов, разработанных на основе собственных оригинальных подходов к анализу. Интеграция алгоритмов основана на использовании общей экспериментальной базы данных (БД), предоставляющей пользователям доступ к данным при помощи графического интерфейса и языка запросов.

АПК способен обрабатывать два принципиально различных вида исходных экспериментальных данных– данные геномных экспериментов и данные протеомных экспериментов. Данные протеомных экспериментов представляют

собой спектры, записанные в формате специфическом для используемого оборудования или стандартизованном формате xml, а данные геномных экспериментов представлены короткими прочтениями (ридами) и соответствующими им значениями качества, длина прочтений варьиует от 1000 нуклеотидов (для ридов, полученных капиллярным сиквенированием) до 50 нуклеотидов (для ридов, полученных на платформа AB Solid 4).

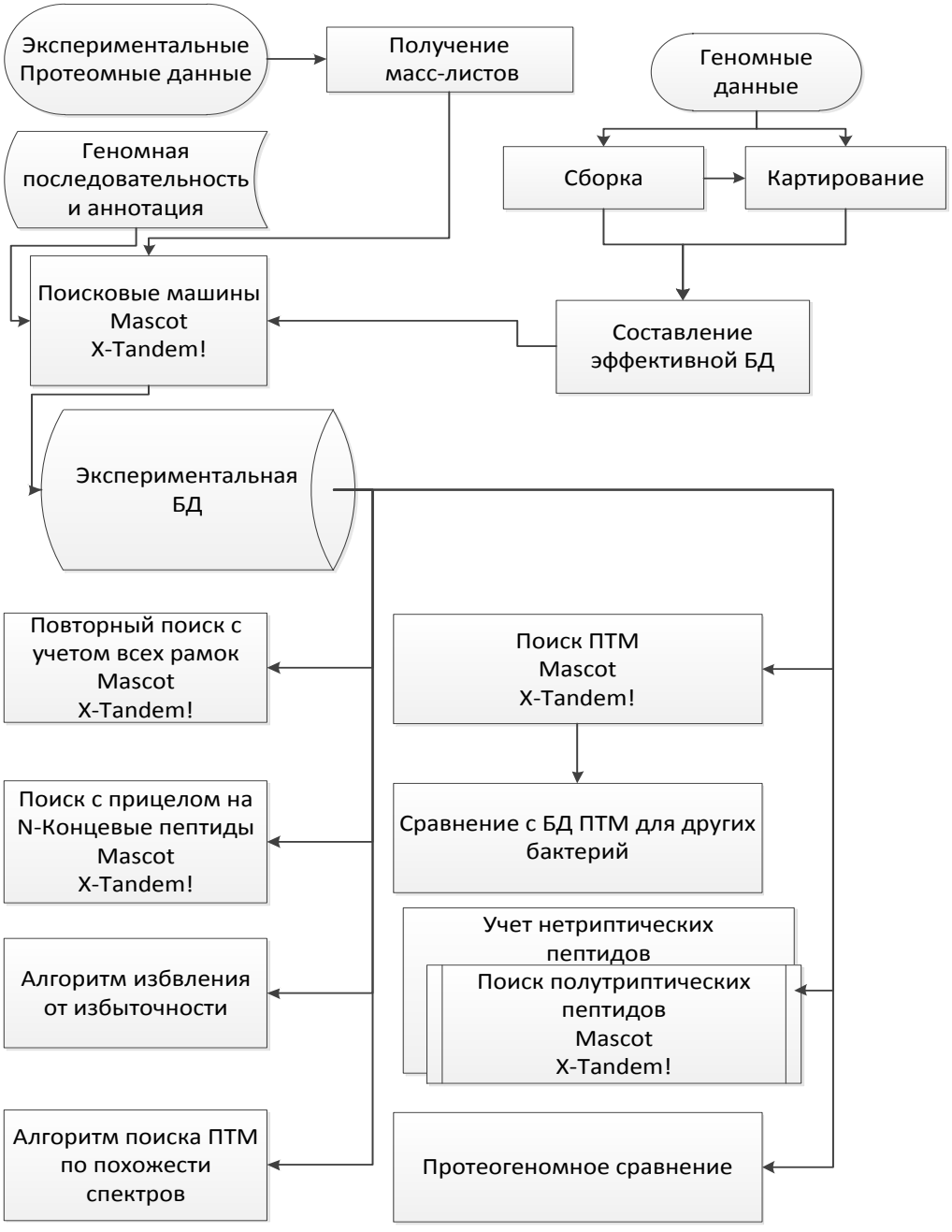

Рисунок 4 Схема автоматизированной системы обработки протеогеномных данных

Среди разработанных алгоритмов стоит выделить несколько алгоритмов, впервые реализованных в виде программного обеспечения – это алгоритм учета неспецифичности трипсина, алгоритм учета геномных данных, алгоритм избавления от избыточности в протоемных идентификациях и алгоритм протеогеномного сравнения.

С использованием алгоритмов была произведена реаннотация геномов, общий результат представлен в сводной таблице 3.

Таблица 3. Статистика по протеомным экспериментам.

Штамм	Экспериментов MS	MS/MS спектров	Обнаружено белков	% от общего	Новых белков	Старт сайт
M.gallisepticum S6	1000	1 млн.	481	68%	0	32
A.laidlawii PG-8A	2000	2 млн.	876	64%	30	40
S.melliferum KC3	1500	2 млн.	521	44%	50	10
D.kamchatkensis 1221n	400	600 тыс.	625	41%	0	10
H.pylory 26695	400	100 тыс	707	47%	0	0
H.pylory J99	200	50 тыс	550	36%	0	0
H.pylori A45	200	50 тыс	604	40%	0	0

Был разработан комплекс подходов, позволяющий работать с лабораторными штаммами и клиническими изолятами, геномы которых не известны. Проведена оценка диапазона межштаммового различия, допустимого для сохранения высокой достоверности протеомных идентификаций, которая для гомоморфных частей геномов может быть выражена следующей формулой:

ОП на белок / 6 (ср. число пеп. на белок) * 2 (идент. пептидов) * % по 2м. пептидам (Формула 1.), где

- ОП на белок – количество однонуклеотидных полиморфизмов на белок,

- 6 (ср. число пеп. на белок) – среднее число пептидов на белок (6 – среднее число),

- 2 (идент. пептидов) – минимальное число пептидов для идентификации белка,

% по 2м. пептидам – процент белков идентифицированный по 2-ум пептидам

Далее была проведена экспериментальная проверка оценки, в двух случаях – работа с лабораторным штаммом *Mycoplasma gallisepticum* S6и клиническим изолятом Helicobacter pylori A45. Полученные результаты позволяют говорить о минимум 90% успешной идентификации, проверка была произведена после получения данных о геномах указанных штаммов. Так же на примере частично последовательности генома была показана возможность учета данных частичного генома для улучшения белковых идентификаций.

Был произведен системный анализ на основе протеогеномного сравнения нескольких трех наборов протеогеномных данных:

1) *Mycoplasma gallisepticum*, *Acholeplasma laidlawii*, *Mycoplasma mobile*
2) *Spiroplasma melliferum*, *Spiroplasma citri*
3) 3 штамма *Helicobacter pylori*: J99, A45, 26695

На основе данных о протеогеномном сравнении *Mycoplasma gallisepticum*, *Acholeplasma laidlawii*, *Mycoplasma mobile,* выращенных на одинаковой среде , выявленное сходство было положено в основу создания представления об общем протеомном ядре– 212 COG. Далее было показано, что на основе имеющихся данных по комплексообразованию у молликут протеомное ядро включает в себя большую часть известных комплексов, при этом среднее количество партнеров у представителей выше, чем среднее по геному. Таким образом, с учетом представлений о пространственной структуре клетки была предложена модель, согласно которой протеомное ядро является основой белкового взаимодействия в клетке, а специфические клеточные функции, разнящиеся от штамма к штамму, имеют единичные контакты с представителями ядра.

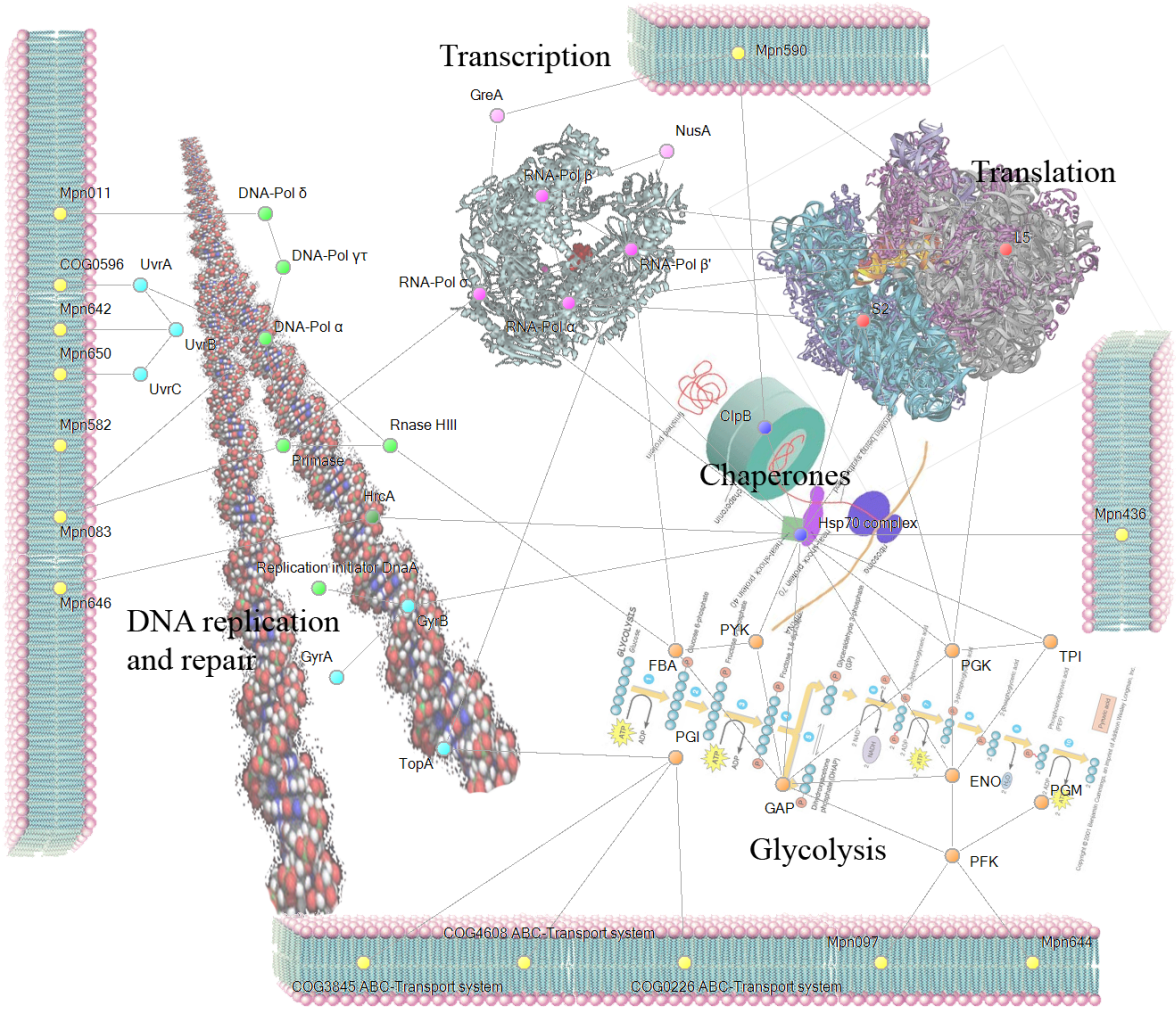

Рисунок 5 Схема взаимодействия белков протеомного ядра находящегося в *Mycoplasma pneumonia*, построенного на основе данных о комплексообразовании.

Протеогеномное сравнение *Spiroplasma melliferum* и *Spiroplasma citri* позволило выявить ряд особенностей в составе протеома *S. melliferum* KC3, которые, по нашему мнению, ассоциированы с высокой патогенностью данного вида для насекомых.

Отличия между протеомами двух видов заключаются в ряде белков, которые в составе генома *S. melliferum* включены в предположительные мобильные элементы (рис. 7). По-видимому, горизонтальный перенос этих белков позволил *S.melliferum* обрести отличительные фенотипические свойства за счет генов экзотоксинов и утилизации хитина, которые определили его патогенность.

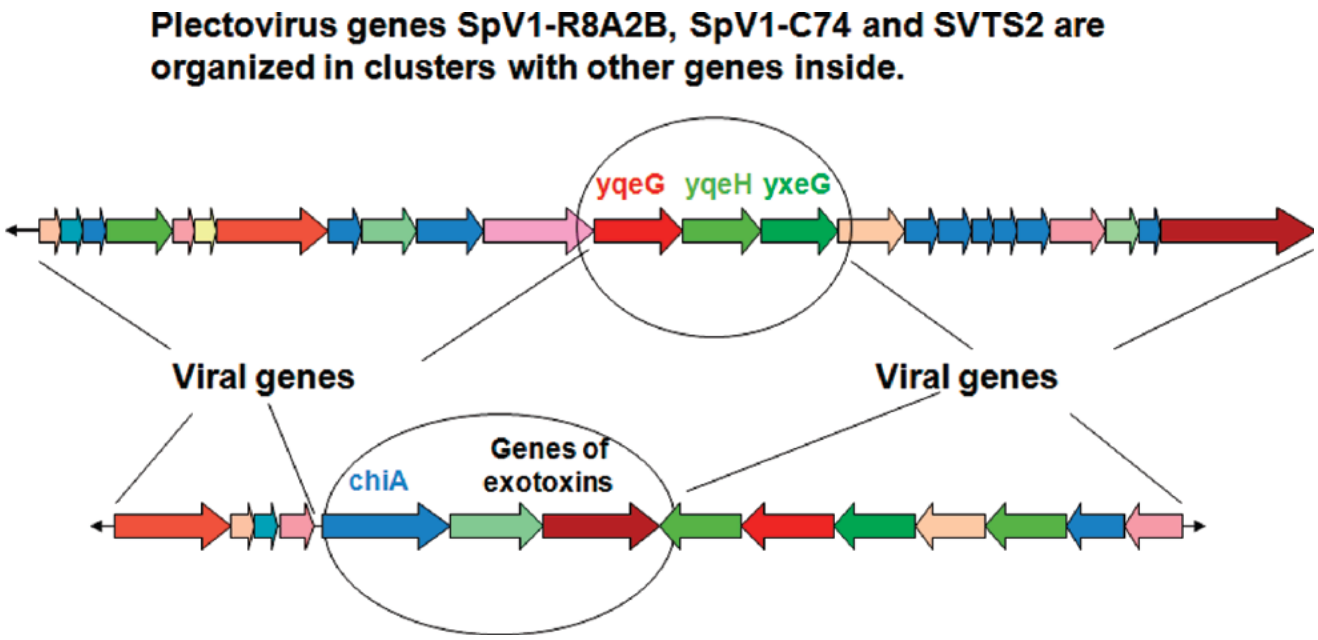

Рисунок 7 Пример организации мобильного островка плектовирусной природы. Вирусные гены окружают смысловые гены и возможно способствуют горизонтальному переносу.

Протеогеномное сравнение 2 лабораторных штаммов и одного клинического изолята *H.pylori* позволило выявить большое разнообразие в белковой представленности для разных штаммов, при этом для клинического изолята наблюдаются уникальные экспрессируемые белки:

Таблица 4. Уникально экспрессирующиеся белки в клиническом изоляте A45.

regulatory protein DniR	связан с повышеным синтезом dissimilatory nitrite reductase
NifU-like protein	индуцируется желчным и желче-кислотным стрессом
hypothetical protein HP1453	
siderophore-mediated iron transport protein (tonB)	фактор вирулентности
hypothetical protein HP0838	
glutamylglutaminyl-tRNA synthetase	гомологична glutamyl-tRNA synthetase hp0476
hypothetical protein HP0573	
cyclopropane fatty acid synthase (cfa)	устойчивость к кислоте и бутанолу
hypothetical protein HP0309	
collagenase (prtC)	Эпителиальный клеточный сиганллинг при возникновении инфекции
lipopolysaccharide 1,2-glucosyltransferase (rfaJ)	Синтез ЛПС
hypothetical protein HP0080	Внешний мембранный
hypothetical protein HP0018	

Часть уникальных белков клинического изолята напрямую или опосредованно ассоциирована с вирулентностью и способностью к трансформации, а о других ничего не известно. Для получения более подробной картины нами предложен оригинальный метод количественного сравнения, основанный на высокой представленности аминокислотных полиморфизмов между геномами хеликобактера.

3.1 Заключение

В ходе выполнения диссертационной работы (2006-2011 гг.) мы столкнулись с необходимостью учета новой информации о механизмах функционирования наиболее просто устроенных микроорганизмов, к числу которых относятся микоплазмы. Появились новые данных о повсеместной представленности некодирующих РНК у бактерий, в том числе и у бактерий с так называемым «сверхплотным» кодированием как, например, у H.pylori, для которых определяется не менее 1000 некодирующих транскриптов. Появились исследования о существенно ранее не описанных биохимических ферментативных комплексах в клетке *Mycoplasma pneumonia* и других бактерий. Появление таких данных обусловило целесообразность и актуальность тщательной реаннотации геномов изучаемых микроорганизмов, для чего нами были использованы протеомные технологии. Реципрокные взаимосвязи между протеомом (комплиментарность, стехиометрия, ПТМ и т.д.) и транскрипционной активностью генома и его структурной вариабельностью предоставляет в настоящее время недостающую основу для синтетической биологии и определения действительной модели жизни.

Выводы

1) Эффективные алгоритмы. использующие данные протеомных экспериментов для протеогеномного профилирования, были разработаны. Алгоритмы позволяют учитывать данные множества поисковых машин, осуществлять поиск ПТМ и N-концевых пептидов, производить реаннтацию геномов и осуществлять проетогеномное сравнение.

2) С использованием алгоритмов были улучшены аннотации геномов Mycoplasma gallisepticum, Acholeplasma laidlawii, Spiroplasma melliferum и Desulfurococcus kamchatkensis.

3) Алгоритмы были использованы для протеогеномного профилирования изолятов и штаммов, для которых геномы не секвенированы или существует только частичная последовательность, была разработана методика оценки достоверности такого профилирования.

4) Алгоритмы позволяют произвести системный анализ и улучшить протеогеномную аннотацию на основе сравнения протеогеномных профилей бактерий и получить биологически релевантные выводы.

Список работ опубликованных по теме диссертации:

1. Alexeev D, Kostrjukova E, Aliper A, Popenko A, Bazaleev N, Tyakht A, Selezneva O, Akopian T, Prichodko E, Kondratov I, Chukin M, Demina I, Galyamina M, Kamashev D, Vanyushkina A, Ladygina V, Levitskii S, Lazarev V, Govorun V. Application of Spiroplasma melliferum Proteogenomic Profiling for the Discovery of Virulence Factors and Pathogenicity Mechanisms in Host-associated Spiroplasmas// Journal of proteome research, 2011. V.11(1) P. 224-36.
2. Lazarev VN, Levitskii SA, Basovskii YI, Chukin MM, Akopian TA, Vereshchagin VV, Kostrjukova ES, Kovaleva GY, Kazanov MD, Malko DB, Vitreschak AG, Sernova NV, Gelfand MS, Demina IA, Serebryakova MV, Galyamina MA, Vtyurin NN, Rogov SI, Alexeev DG, Ladygina VG, Govorun VM. Complete Genome and Proteome of Acholeplasma laidlawii// Journal of bacteriology, 2011. V.193(18), P.4943-53.
3. Fisunov GY, Alexeev DG, Bazaleev NA, Ladygina VG, Galyamina MA, Kondratov IG, Zhukova NA, Serebryakova MV, Demina IA, Govorun VM. Core proteome of the minimal cell: comparative proteomics of three mollicute species// PloS one, 2011 V. 6(7).
4. Дёмина И.А., Серебрякова М.В., Ладыгина В.Г., Галямина М.А., Жукова Н.А., Алексеев Д.Г., Фисунов Г.Ю., Говорун В.М. Сравнительная протеомная характеристика микоплазм (молликут)// Биоорганическая химия. 2011. Т. 37. № 1. С. 70-80.
5. Momynaliev, K. T., Kashin, S. V., Chelysheva, V. V., Selezneva, O. V., Demina, I. A., Serebryakova, M. V., Alexeev, D. Functional divergence of Helicobacter pylori related to early gastric cancer// Journal of proteome research, 2010. V. 9(1), P. 254-67.
6. Alexeev, Bazaleev, Govorun Semantic relationships derived from experemental analysis experience help to proceed and visualize experimental data. Proceedings of International Conference on Bioinformatics of Genome Regulation and Structure (BGRS' 2010) Novosibirsk, Russia, June 20—27, 2010. P.25
7. Altukhov, Ischenko, Alexeev, Bazaleev, Uvarovskiy, Tyakht Web-application for comparative structural and functional analysis ofprokaryotic genomes sequencing

data. Proceedings of Moscow Conference on Computational Molecular Biology Moscow, RussiaJuly 21–24, 2011, P.36.